AD-A119 138 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
COMPUTER-GENERATED INDICES OF STUDENT PERFORMANCE.(U)
AUG 82 K JOHNSON
UNCLASSIFIED NPRDC-SR-82-34 NL

1 OF 1
AD A
119138

END
DATE
FILMED
10-82
DTIC

NPRDC SR 82-34

AUGUST 1982

COMPUTER-GENERATED INDICES OF STUDENT PERFORMANCE

NAVY PERSONNEL RESEARCH
AND
DEVELOPMENT CENTER
San Diego, California 92152

NPRDC Special Report 82-34 August 1982

COMPUTER-GENERATED INDICES OF STUDENT PERFORMANCE

Kirk Johnson

Reviewed by
J. S. McMichael

Released by
James F. Kelly, Jr.
Commanding Officer

Navy Personnel Research and Development Center
San Diego, California 92152

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPRDC SR 82-34
2. GOVT ACCESSION NO.: AD-A119 138
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): COMPUTER-GENERATED INDICES OF STUDENT PERFORMANCE
5. TYPE OF REPORT & PERIOD COVERED: Final Report FY 79
6. PERFORMING ORG. REPORT NUMBER: 14-81-8
7. AUTHOR(s): Kirk Johnson
8. CONTRACT OR GRANT NUMBER(s):
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Navy Personnel Research and Development Center, San Diego, California 92152
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: ZF55.522.002, Z1176-PN.01
11. CONTROLLING OFFICE NAME AND ADDRESS: Navy Personnel Research and Development Center, San Diego, California 92152
12. REPORT DATE: August 1982
13. NUMBER OF PAGES: 40
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (Continue on reverse side if necessary and identify by block number):
Individualized instruction
Self-paced instruction
Computer-managed instruction
Computer-based instruction
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

A number of indices that can be generated by computers as aids in managing students were evaluated by applying them to historical data on student performance drawn from the Navy's Aviation Fundamentals Course. It was found that several of the indices now in use are less powerful or less desirable than available alternatives. Some of these alternative indices could be adopted with little difficulty; others would require fairly substantial modifications of the existing system.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

FOREWORD

This research and development was conducted in support of exploratory development task area ZF55.522.002 (Methodology for Development and Evaluation of Navy Training Programs) and advanced development subproject Z1176-PN.01 (Improving the Navy's Computer-managed Training System) under the sponsorship of the Deputy Chief of Naval Operations (Manpower, Personnel, and Training) (OP-01).

The information provided in this report should be useful to designers of computer-based instructional systems.

JAMES F. KELLY, JR.
Commanding Officer

JAMES J. REGAN
Technical Director

SUMMARY

Problem

Individually-paced instruction is effective in reducing the time required to reach training objectives, but it complicates the task facing the instructors. They must monitor a large and continually changing population of students who differ widely in both ability and motivation, and who are working at a variety of points within an extended series of heterogeneous training modules. In spite of this complexity, the instructor must decide who is most in need of assistance, who should receive incentives or disincentives, and who should be dropped from the course.

Purpose

The purpose of this investigation was to evaluate alternative indices of student performance by applying them to historical data drawn from an individually-paced Navy course. The focus was on the major indices currently provided by the Navy's computer-managed instruction (CMI) system and on several of the more promising alternatives to each.

Method

Performance data were collected on a large sample of students in the aviation fundamentals course, a short course representative of Navy courses taught by means of CMI. This sample was split into two subsamples. The first of these was used to develop indices relevant to major decisions that must be made by the instructor. These indices included most of those provided by the current system plus several major alternatives. The ways in which the indices actually function were then evaluated by applying them to data from the second subsample.

Much of the analysis was devoted to the comparison of alternative techniques for predicting student performance since these predictions provide standards against which to evaluate actual performance. These predictions were evaluated in terms of both relative (i.e., normative or correlational) and absolute accuracy. Procedures designed to select students for various forms of attention (e.g., as candidates for incentives) were evaluated in terms of their agreement with alternative procedures and their susceptibility to various forms of bias.

Results

In general, information about past performance contributed substantially to the accurate prediction of future performance. The technique of adjusting initial predictions on the basis of past performance (which is used in the present system) was slightly inferior to alternative techniques in terms of relative accuracy and was substantially less accurate than the other techniques as a means for predicting actual times. In fact, there were situations in which these adjusted predictions were less accurate than the original predictions. Predictions using the sum of times on previous modules were as accurate as predictions using times on individual modules.

A number of specific weaknesses in the existing CMI system were identified:

1. General criteria for defining unexpectedly poor performance on individual modules tended to focus too much attention on short modules.

2. When predictions were based on raw data, criteria defining unexpectedly poor performance on individual modules tended to focus too much attention on the brighter students.

3. Variations in indices based on cumulative time were too dependent on position in course to be useful in detecting local variations in performance. The existing system tended to focus too much attention on the early portions of the course.

4. Procedures used for the allocation of both incentives and disincentives were not sufficiently sensitive to recent variations in performance.

5. The system for awarding incentives, which is based directly on deviations from initial predictions, was biased against the brighter students.

6. The system for assigning students to night school, which is based on a ratio between deviations from predicted performance and time remaining in the course, tended to concentrate too many assignments in the latter portions of the course.

Conclusions

This project identified a number of procedures being used in the present system that are less powerful or less desirable than alternative procedures. Some of the alternative procedures could be adopted with little difficulty. Others would require slightly more complex modifications. Still others would require major modifications.

Recommendations

1. Certain procedures used in the present system could be improved with minimum cost. Among these are those used in displaying performance on recent modules in selecting students for night school and in selecting students for certain positive incentives. The CMI system should be modified to incorporate these new procedures.

2. Other procedures could be improved only through revisions that are fairly extensive or that might place serious limitations on other aspects of the system. Among these potential revisions are new techniques for basing predictions on past performance, the use of individual rather than general criteria for various types of selection, and the use of predictions based on transformed data. The costs and benefits associated with procedures of this kind should be analyzed, and decisions should be made about their net utility. If it is decided that revisions should be made, an additional choice should be made between immediate implementation and postponement until there is a more general revision of the CMI system as a whole.

CONTENTS

LIST OF TABLES

Page

INTRODUCTION

Problem

Individually-paced instruction is effective in reducing the time required to reach training objectives, but it complicates the task facing the instructors. They must monitor a large and continually changing population of students who differ widely in both ability and motivation, and who are working at a variety of points within an extended series of heterogeneous training modules. In spite of this complexity, the instructor must decide who is most in need of assistance, who should receive incentives or disincentives, and who should be dropped from the course.

Purpose

The purpose of this investigation was to evaluate alternative indices of student performance by applying them to historical data drawn from an individually-paced Navy course. The focus was on the major indices that are currently provided by the Navy's computer-managed instruction (CMI) system and on several of the more promising alternatives to each.

Background

The Navy's CMI system collects and stores a great deal of information about student performance. To be useful, this raw data should be reduced to a relatively small number of indices that (1) are based on accurate summaries of performance, (2) require a minimum of additional processing by the instructor, (3) are oriented as precisely as possible toward particular decisions about student management, and (4) are not distorted by irrelevant factors.

CMI provides reports that are designed to help the instructor manage students more effectively. These reports contain a variety of indices oriented toward various aspects of student performance; each index is designed to assist the instructors with one or more of the decisions they must make in the management of individual students.

The reports have evolved in a somewhat haphazard manner over a period of years. At no time has there been a systematic evaluation of the ways in which various indices actually function when applied to performance of the kind typically found in Navy courses, nor have there been systematic comparisons between existing indices and alternatives that might improve the system.

A glossary of abbreviations and acronyms used frequently in this report is provided (pp. 31 and 32).

APPROACH

Aviation Fundamentals (AFUN) Course

Data were drawn from the aviation fundamentals (AFUN) course, which provides an introduction to topics such as aircraft systems, aircraft handling, maintenance documentation, and hand tools. The course is individually-paced, computer-managed, and lasts an average of about 62 hours. It is organized into 17 regular instructional modules, 1 shop module, and 4 special test modules.

The typical instructional module is divided into two or three lessons taught by means of programmed instruction. Multiple-choice end-of-module tests are graded by the computer. Students failing more than 10 percent of the questions on a given lesson are told to restudy the lesson and are then given a test covering only that lesson. A second failure is followed by the assignment of another lesson test. A student failing for a third time is referred to the instructor who, after evaluating the situation, might clear the student from the module, assign still another test, or initiate procedures for dropping the student from the course.

A special test module is assigned after every four or five instructional modules. It covers all lessons taught since the last special test module (though not exhaustively). In cases of failure, lesson tests are assigned in the same way that they were assigned following failure on a module test.

Sample of Modules

Analyses were limited to a randomly selected sample of ten modules--Nos. 1, 2, 4, 6, 8, 10, 12, 14, 19, and 20. All were regular instructional modules except for 6, which was a special test module.

Sample of Students

The sample consisted of 715 Navy students who had graduated from the course over a period of approximately 2 months in 1978. Half of the sample was randomly assigned to a development sample; and the other half, to a cross-validation sample.

Predicted Times

Most indices were based on predictions of the time that individual students should need to complete specific segments of the course. Training time has intrinsic importance as a basis for various administrative actions (e.g., requests for orders) and it is probably the single most important element in training cost. In individually-paced courses that require high levels of mastery on most training objectives, it also provides one of the most sensitive indices of student ability, effort, and effectiveness.

The types of predictors are described in the following paragraphs. All except No. 6 (Fading of Weights) (FADE) were used to predict times on individual modules and groups of consecutive modules. FADE was used only to predict times for groups of consecutive modules.

1. Aptitude (APT)

The basic predictions used in the existing system are developed from multiple regression equations in which year of birth, years of educations, and nine scores from the Armed Services Vocational Aptitude Battery (ASVAB) are used as predictors. Predictions of this kind will be referred to as APT (for aptitude).

2. Individual Module Times (IND)

Most of the predictions considered in this investigation were developed by adding information about the student's past performance to the information provided by the predictors used in APT. IND was developed by adding the times for each individual module that the student had already completed to the set of predictors used to develop APT.

3. Sum of Module Times (SUM)

Another type of prediction was developed by adding the sum of time on all previously completed modules to the set of predictors used for APT. The use of the sum instead of individual module times simplifies the equations, particularly in long courses that contain many modules. However, the simplification is at the expense of any increase in predictive power that might be achieved through the differential weighting of modules that are more (or less) powerful predictors than other modules.

4. Adjusted Predictions (ADJ)

ADJ, which is used for certain applications in the existing system, was developed by multiplying an initial prediction based on APT by an adjustment or correction factor consisting of the ratio of actual time on all preceding modules to predicted time (also based on APT) on all preceding modules. In other words, any systematic tendency toward over- or underestimation in previous portions of the course is used to compensate for similar tendencies that might affect predictions on present or future portions of the course. These adjustments may sound somewhat complicated, but they provide two types of predictions: one based only on pretraining variables and the other on performance in earlier parts of the course as well. Both are from the same set of basic equations.

Each type of prediction described in the preceding paragraphs can be used as the basis for predicting a variety of values (e.g., individual module times, the times required for course completion from various points in the course, and total training time in the course). All of these values are predicted and used in the existing system. Since the ways in which the various types of predictions function may differ for different kinds of predicted values, a variety of comparisons was required.

There are two additional types of predictions that have more limited applications and that were evaluated in fewer situations: past performance and fading of weights.

5. Past Performance (PAST)

ADJ uses the relative error between prediction and actual performance to adjust another prediction. An obvious alternative is to substitute group predictions (i.e., means) for the individual predictions used in ADJ. The predictions based on mean times will differ from the predictions based on APT only to the extent that there are student-by-module interactions in APT. Since individual differences in predictions based on mean times are determined solely by individual differences in the students' past performance, predictions of this kind will be referred to as PAST.

6. Fading of Weights (FADE)

FADE is a variation on APT that is differentially oriented toward the prediction of recent performance. Equations were developed to predict weighted sums of times on previous modules. These sums were formed by multiplying each weighted sum for previous modules by .8 before adding the time for a new module. As a result, the weights for previous modules decreased with increasing distance from the point of prediction.

Relative and Absolute Accuracies of Predicted Times

Predictions were evaluated in terms of both relative and absolute accuracy. Relative accuracies (Rel.) were measured by the coefficient of determination or the percentage of variance in times accounted for by the best linear transformation of predicted times.

Stated somewhat differently, this coefficient is one minus the square of the ratio of the standard error of estimate to the standard deviation of the scores.

Absolute accuracies (Abs.) were measured by substituting the root mean square (RMS) of the differences between actual and predicted scores for the standard error of estimate.

When predictions are based on multiple regressions, the indices of relative and absolute accuracy will be the same in the development sample. This will not be true when predictions are based on other techniques (e.g., ADJ), nor will it be true in the cross-validation sample. The indices of relative accuracy will never be negative, but the indices of absolute accuracy will be negative whenever the RMS error is larger than the standard deviation. This might occur because the sign of the slope has changed, because the means are different, or because the predictions are too variable (over-prediction). When it does occur, it indicates that the mean of the sample would have provided a more accurate prediction than the predicted scores being evaluated.

Selection of Students

There are situations in which a prediction of time is used directly as the basis for a decision, but more frequently it is used only as a standard for evaluating student performance. In these cases, the prediction of time is generally incorporated into an index that either triggers some action by the computer or is oriented as directly as possible toward some specific decision by the instructor.

In most cases, the index will be some function of the discrepancy between actual and predicted times. When this index exceeds some predetermined cutting score, the student will be selected or flagged by the computer. Since indices of this kind represent a major focus of the evaluation, a preliminary summary of their two principal uses may be helpful.

One use is to detect students who are having unexpected or unpredictable difficulties. Such difficulties are hard to detect, since a bright student can be working much more rapidly than a less able student but still be working at a level far below that at which he or she should be working. Once these unexpected difficulties are detected, however, they are probably easier to correct than are the more predictable difficulties, since, by definition, they have been avoided by most students of comparable ability.

The second use of the indices is to detect students who are devoting significantly more or less effort to their studies than are other students. When ability is partialled out of measured performance, it leaves a residual that can be interpreted as an index of effort or motivation. A system that allocates incentives on the basis of such an index is more equitable and is probably more effective than a system that allocates them on the basis of performance alone.

In order to facilitate comparisons between alternative techniques, data from the development sample were used to set cutting scores at values that would select roughly 15 percent of the students. This is an arbitrary parameter, but it is fairly representative of the percentages that might be used for purposes such as assigning students to night school or flagging students who are experiencing unexpected difficulties.

Evaluation of Selection

The greatest danger with relatively automatic selections of this kind is that they will be biased in unintended ways; that is, they will select too many or too few students

who have certain characteristics or who are in certain parts of the course. Tests, consisting of both categorical comparisons and correlations, were made to determine how frequency of selection varied as a function of factors such as student aptitude, various module characteristics, and the student's position in the course.

Additional tests were made to determine the extent to which the same students would be selected by a given index at different points in the course and by different indices at the same points in the course. In order to keep these tests independent of differences in the frequency of selection and to avoid extensive reanalysis of the data, a somewhat indirect index of agreement was used. Tetrachoric correlation coefficients were estimated from the ratio of diagonal products using a published table.[1] This same table was then used as the basis for estimating the percentage of overlap among selected students that would have been found had the frequency of selection been the same in all groups.

Preliminary Analyses

Several preliminary analyses were done prior to the actual comparisons of alternative indices.

Distribution of Predictors

ASVAB scores, as expected, were distributed quite normally (median kurtosis = .5; median skewness = .0). Years of education were more skewed towards low values (kurtosis = 4.9; skewness = -.7) than any of the ASVAB scores, but the difference was not large. Year of birth, however, was markedly skewed toward the earlier dates or older ages (kurtosis = 13.9; skewness = -1.3). A logarithmic transformation ($\log_{10}$ (year of birth-1962)) served to normalize the distribution (kurtosis = .5; skewness = -.1), and these transformed values were used as the actual predictors.

Distribution of Times

Similar analyses were done for times on each of the 10 modules and for cumulative times at the completion of each of these modules. The times for individual modules were all skewed toward long times (median kurtosis = 25.7; median skewness = 4.1). The cumulative times were also skewed (median kurtosis = 7.6; median skewness = 2.0) but to a lesser degree. Most of the distributions appeared to be reasonably continuous; there were few clear breaks or irregularities that could be used to identify erroneous times at either end of the distribution.

It was found that a logarithmic transformation ($\log_{10}$ (X = .1) + 1.0) served to normalize both individual times (median kurtosis = .7; median skewness = .4) and cumulative times (median kurtosis = .5; median skewness = .2). In most cases, the change was from a distribution that might be described as markedly asymmetrical to one that appeared to be reasonably normal. The transformation also served to increase the average correlation[2] between individual module times and predictors from -.10 to -.14, and average correlations between module times from .22 to .27.

---

[1] Davidoff, M. D., & Goheen, H. W. A table for the rapid determination of the tetrachoric correlation coefficient. Psychometrica, 1953, 18, 115-121.

[2] All average correlations reported in this evaluation were computed using Fisher's *z* transformation.

The higher correlations and more regular distributions represent advantages for the transformed times, but there are also disadvantages which will be discussed later. It was decided, therefore, to analyze indices based on both kinds of data.

Comparison with Another Course

Since the correlations found in these preliminary analyses seemed low, they were compared with those found in the Basic Electricity and Electronics (BE&E) Course.[3] The average correlation of module times with ASVAB scores was somewhat higher in BE&E (-.17, as opposed to -.12), and the average correlation between module times was substantially higher (.41, as opposed to .22).

Editing

There are a number of ways in which erroneous times can be recorded by the computer. A failure to make the proper correction for such errors will create times that are either too long or too short. In the current system, the data to be used as the basis for predictions are edited by removing values that, in the opinion of experienced instructors, are probably erroneous.

Two approaches to editing were investigated. The first was editing on deviations from the mean of each distribution, an algorithmic version of the technique that is currently being used. The second was editing on deviations from a preliminary regression line based on IND. The latter technique considers both aptitude and previous performance in the definition of times that are "probably erroneous" and can detect such times even when they fall within an otherwise reasonable range. Both techniques were investigated at two levels of editing: One removed 2.5 percent of the cases from each end of the distribution and the other removed 10.0 percent of the cases from each end of the distribution.

Edited data from the development sample were used to derive equations that were then applied to the unedited cross-validation (C.-V.) sample. Separate analyses were made for APT, IND, and SUM, using both untransformed and transformed times. Since this entailed 30 analyses for each module, the evaluation was limited to a randomly selected set of four modules (4, 8, 12, and 20). The data are given in Table 1.

Editing tended to have its expected effects in the development sample. Editing around the mean tended to decrease coefficients, due to the restriction of range. Editing around the preliminary regression line tended to increase coefficients, due to the elimination of the less predictable cases. In each case the effect was increased at the higher level of editing. However, the effects of editing around the mean were fairly small, suggesting a disproportionately large reduction in error variance. This was supported by an analysis indicating that the shrinkage predicted from the reduction in variance was substantially larger than the shrinkage that was actually found.

The effects of editing in the cross-validation sample tended to differ as a function of transformation. For untransformed data the effects were slightly beneficial, and for transformed data they were slightly detrimental. These effects were somewhat more consistent for relative accuracy than for absolute accuracy. On the average, there was

---

[3]Federico, P-A., & Landis, D. B. Predicting student performance in a computer-managed course using measures of cognitive styles, abilities, and aptitudes (NPRDC Tech. Rep. 79-30). San Diego: Navy Personnel Research and Development Center, August 1979. (AD-A070 748)

Table 1

Effects of Editing: Percentage of Variance Accounted for by Different Kinds of Prediction in the Development and Cross-validation Samples

Prediction	No Editing	Editing Deviations from the Means		Editing Deviations from Prelim. Regression Line	
		2.5%	10.0%	2.5%	10.0%
APT					
Dev.	9.9	9.6	10.7	13.7	18.7
C.-V. Rel.	4.6	5.5	5.3	5.4	5.7
C.-V. Abs.	3.6	4.7	1.5	4.4	2.2
IND					
Dev.	28.0	23.9	23.7	43.3	49.2
C.-V. Rel.	12.7	13.6	13.9	13.9	14.9
C.-V. Abs.	7.4	12.5	9.3	10.2	11.3
SUM					
Dev.	20.8	19.5	18.6	31.6	40.1
C.-V. Rel.	14.5	15.2	15.6	15.4	16.4
C.-V. Abs.	11.9	14.5	9.6	12.8	13.9
APT Trans.					
Dev.	12.5	11.5	10.8	16.4	23.2
C.-V. Rel.	10.9	10.4	9.5	10.7	10.4
C.-V. Abs.	10.9	10.3	8.5	10.4	9.1
IND Trans.					
Dev.	29.6	28.1	24.3	40.3	53.7
C.-V. Rel.	21.1	20.7	19.4	20.7	21.1
C.-V. Abs.	19.0	19.7	17.7	16.8	18.5
SUM Trans.					
Dev.	24.4	24.0	20.3	33.5	46.7
C.-V. Rel.	20.6	20.2	19.9	20.4	20.4
C.-V. Abs.	19.3	20.1	17.8	18.8	18.8

little difference between the two editing procedures. The lower level of editing around the mean tended to be the most beneficial procedure and the higher level of editing around the mean tended to be the most detrimental procedure.

The major difficulty with the foregoing evaluation is that, if errors are removed from the development sample and not from the cross-validation sample, an equation based on the development sample will tend to over-predict in the cross-validations sample. This will tend to inflate the overall mean square error in the cross-validation sample, even

though the mean square error for valid, nonerroneous times would be reduced. Unfortunately, it is impossible to test the extent of such an effect without some independent means for the identification of errors.

In summary, algorithmic editing of the kind investigated here does not seem to have a large effect on the accuracy of predictions. A low level of editing would probably be somewhat beneficial and would provide protection against the potential effects of a few large errors but, in the interest of simplicity, the remainder of the analyses were performed on unedited data.

RESULTS

Prediction of Individual Module Times

In the current system, a prediction of the time required by an individual student to complete a particular module is provided on the individual's lesson guide when the module is assigned, again when it is completed, and again on the daily student progress report (DSPR). These predictions can be used as a basis for automatically flagging, on the DSPR, students who have exceeded a reasonable time in the assigned module. They also serve as the basis for predicting cumulative times since, with untransformed data, regression equations are additive just as the times themselves are additive.

Accuracy of Predictions

The predictive accuracies of APT, IND, SUM, and ADJ were averaged over the sample of 10 modules. Separate averages were computed for the development sample and for both relative and absolute accuracies in the cross-validation sample. The data for ADJ in the development sample are absolute accuracies. Separate averages were also computed for untransformed and transformed times. For transformed times, the arithmetic used in computing ADJ was that appropriate for a logarithmic variable (i.e., the ratio was computed by subtraction and the correction by addition). The resulting percentages of variance are given in Table 2.

Table 2

Percentage of Variance Accounted for by Different Kinds of Prediction: Individual Modules

	Untransformed				Transformed			
Sample	APT	IND	SUM	ADJ	APT	IND	SUM	ADJ
Dev.	9.3	23.9	18.8	12.5	14.2	27.9	23.5	17.3
C.-V. Rel.	5.5	13.1	14.0	14.4	10.8	19.4	19.3	19.3
C.-V. Abs.	3.2	7.3	10.0	7.4	9.4	16.9	17.8	10.4

The accuracies of predictions that take past performance into account (IND, SUM, and ADJ) were substantially better than those of predictions that do not (APT). Differences in the relative accuracies of the former were slight, whether transformed or

not. For untransformed data, SUM had a greater absolute accuracy than did either IND or ADJ. For transformed data, both SUM and IND had greater absolute accuracies than did ADJ. There was considerable variability over modules. It was not uncommon to find negative values for absolute accuracy.

PAST (not shown in Table 2) was compared to ADJ using untransformed data. PAST was found to be slightly better than ADJ in terms of both relative (.146 vs. .144) and absolute (.093 vs. .074) accuracy. Table 2 shows that transformation increased both the relative and absolute accuracies of all four predictors. The relative improvement was greatest for APT.

One of the major difficulties with predictions based on transformed data occurs when either the student or the instructor must make a direct interpretation of the predicted time. The predicted module times printed on the lesson guides, for example, provide students with reasonable goals and help them to evaluate their efficiency in reaching those goals, but a prediction of log time would have little value in such a situation. It is possible, however, to make a prediction based on transformed data and then to reverse the transformation before displaying the prediction. The effects of such reversals were investigated for both APT and SUM. In each case, the accuracies following the reversal were substantially lower than those prior to the reversal. In fact, the accuracies of the reversed predictions tended to be quite close to those of predictions based on untransformed data. For APT, the relative accuracy of reversed predictions was .063 as opposed to .055 for predictions based on untransformed data; the absolute accuracies were .021 and .032 respectively. For SUM, the relative accuracy of reversed predictions was .149 as opposed to .140 for untransformed data; the absolute accuracies were .120 and .100 respectively.

The predictions provided by IND and SUM should be fairly similar in most cases. Since SUM is much easier to use and enjoyed a slight advantage in terms of accuracy, IND was dropped from further consideration.

Selections

Indices based on module predictions are generally used to identify students who are not performing as well as might be expected, so the following analyses focus on positive deviations from predicted times.

Most selections in the present system are triggered when the ratio between actual and predicted times exceeds a fixed cutting score. An obvious alternative would be a trigger based directly on the size of the deviation. The ratio is used under the assumption that it will compensate for certain forms of bias that commonly result from time score distributions. Two preliminary analyses were designed to test this assumption. In both cases, predictions were based on APT using untransformed data.

When a fixed deviation was used as the cutting score, the percentage of students selected varied considerably across modules ($\sigma = 5.87$) and was highly correlated with mean module times (.93). The use of a fixed ratio reduced both the variability ($\sigma = 2.37$) and the correlation (-.59). Estimates based on the regression lines indicated that, with a fixed deviation, the likelihood of selection from a 6-hour module would be over 400 percent greater than that from a 1-hour module; with a fixed ratio, the likelihood of selection from a 1-hour module would be about 30 percent greater than that from a 6-hour module.

The variance of standardized residuals was much larger for students in the upper thirds of the predicted time distributions (1.58) than for students in the lower thirds (.46).

It was assumed that, if simple deviations were used as cutting scores, selection would be more likely for students with long predicted times. However, an examination of actual selections indicated a near-equal likelihood of selection from each third of the predicted time distributions (14.4, 15.7, and 14.8% for the high, middle, and low thirds respectively). The use of ratios as cutting scores had the anticipated effect of shifting selections toward students with short predicted times (11.4, 13.9, and 18.7% for the high, middle, and low thirds respectively).

Since the transformations were effective in normalizing the time distributions, simple deviations were considered for use as cutting scores with predictions based on transformed data. A preliminary analysis indicated that the percentage of students selected differed considerably from module to module (standard deviations of 4.03, 4.48, and 4.81 for APT, SUM, and ADJ, respectively). Rather than shift to completely different cutting scores for each module, a compromise, consisting of a constant multiplier for the standard error of estimate (k x SEE), was used. This minimized both differences among modules (standard deviations of .87, 1.41, and 2.00 for APT, SUM, and ADJ respectively) and correlations with various parameters of the module time distributions.

Cutting scores that selected 15 percent of the students from the development sample tended to be quite similar for the various types of predictions, in spite of the sizeable differences in the accuracies of the predictions. For APT, SUM, and ADJ, the ratios used for untransformed data were 1.52, 1.52, and 1.54 respectively; the simple differences used for transformed data were .24, .22, and .23 respectively; and the multipliers used for the standard errors of estimate, again for transformed data, were .99, .97, and .98 respectively.

Table 3 indicates the variability of selections from module to module in the cross-validation sample for (1) general cutting scores, applicable to all modules and chosen so as to select 15 percent of the students from the development sample as a whole, and (2) individual cutting scores, chosen separately for each individual module in the development sample so as to select exactly 15 percent of the students from that module. It should be noted that individual cutting scores for k x SEE are exactly the same as the individual cutting scores for simple deviations. In general, the individual cutting scores were not much better than the general cutting scores in minimizing the variability among modules.

Table 3

Standard Deviations Across all Modules of Percentage Selected for Unexpectedly Poor Performance by Different Indices and Cutting Scores: Individual Modules

Type of Prediction	Type of Data	Basis for Cutting Score	Type of Cutting Scores	
			General	Individual
APT	Untransformed	Ratio	2.69	1.27
	Transformed	k x SEE	1.78	1.48
SUM	Untransformed	Ratio	2.81	2.13
	Transformed	k x SEE	1.25	1.65
ADJ	Untransformed	Ratio	3.11	2.53
	Transformed	k x SEE	1.51	2.50

The data in Table 4 indicates the way in which selection within modules varied as a function of predicted time. For predictions based on untransformed data, there was a tendency for selection to increase with decreases in predicted time. This tendency was worse for ADJ than for APT or SUM. The use of predictions based on transformed data eliminated this tendency for both APT and SUM, but not for ADJ. For ADJ, the likelihood of selection in the lower thirds of the predicted time distributions was over twice as great as it was in the upper thirds.

Table 4

Percentage Selected by Different Indices of Unexpectedly Poor Performance as a Function of Predicted Time: Individual Modules

Predicted Time	APT	APT Trans.	SUM	SUM Trans.	ADJ	ADJ Trans.
Long	11.4	13.1	11.8	13.9	9.7	10.9
Medium	13.9	16.0	13.1	15.1	14.3	14.7
Short	18.7	15.7	19.3	15.9	22.7	22.3

Overlap of Students Selected by Various Methods

Table 5 gives the estimated overlap of students selected by various methods. The Time variable represents selections based directly on long module times (no predictions) and was included as a baseline for the other techniques. APT was more closely related to Time than were either SUM or ADJ, but this may reflect no more than the lower accuracy of APT. The two techniques that take past performance into account (SUM and ADJ) were more closely related to each other than to APT. The overlaps between transformed and untransformed indices were fairly high.

An estimated overlap of only 80 percent was found between ADJ and PAST for untransformed data. This indicates that the student-by-module interactions in APT did make a difference, but that these interactions did not contribute to the accuracy of predictions.

Table 6 indicates how much selections overlapped across modules. The average overlaps for the two techniques that take past performance into account (SUM and ADJ) were at the chance level (15%). The overlaps for Time and APT were not high, reflecting the low correlations between modules cited at the outset. For all techniques except ADJ, there was a tendency for overlaps with the nearer modules to be greater than overlaps with more remote modules.

The distributions of total selections over students were determined largely by the average overlap between modules. The distribution for Time and APT had a modal selection of zero, a few cases in which students were selected seven or eight times, a median of about 1.00, a standard deviation of about 1.65, and indices of kurtosis and skewness of about 3.15 and 1.50 respectively. The distributions for SUM and ADJ resembled the distribution for random selection and all had a modal selection of one, no cases in which students were selected over six times, a median of about 1.37, a standard deviation of about 1.15, and indices of kurtosis and skewness of about .20 and .65 respectively.

Table 5

Percentage of Overlap of Students Selected by Different Indices of Unexpectedly Poor Performance: Individual Modules

Basis of Selection	APT	SUM	ADJ
Untransformed			
Time (long)	77	61	56
APT	--	73	69
SUM	--	--	88
Transformed			
Time (long)	85	69	57
APT	--	80	64
SUM	--	--	86
Untransformed with Transformed	88	85	88

Note. These values were computed from the last seven modules in the sample, since relationships early in the series are not representative of an extended series.

Table 6

Percentage of Overlap of Students Selected for Unexpectedly Poor Performance on Different Individual Modules

Type of Data	Average	Distance (Intervening Modules) Between Modules[a] 1	3	5
Untransformed				
Time (long)	28	30	28	27
APT	26	32	25	24
SUM	15	20	13	14
ADJ	15	16	16	15
Transformed				
APT	25	28	25	26
SUM	16	22	16	16
ADJ	14	15	16	15

[a]These values were computed between modules in the sample of ten; there was generally one module between each of the sampled modules, as suggested by these headings, but this was not always the case.

Groups of Modules

The analysis of predictions for groups of modules and of the indices based on them is complicated by the variety of groups for which predictions might be made and by the fact that a single decision, for example, the selection of students for night school, might be based on indices developed from several different groups of modules. Because of this complexity, separate sections will be devoted to the accuracy of predictions for several different groups. These will be followed by sections that focus on different applications and the kinds of selections provided by alternative indices. Discussions of use will be incorporated into the individual sections.

Accuracy: Cumulative Times (CUM)

Predictions of cumulative times, other than those for the total course, are generally made for modules that have already been completed and are used only for purposes of comparison with actual performance. Past performance would not play a part in such predictions.

Table 7 indicates the average accuracies of CUM APT for the cumulative times calculated at the end of each of the ten modules. There was a considerable advantage for predictions based on transformed data, just as there was for individual modules. There was far less shrinkage in moving from the development to the cross-validation sample than there was for the individual modules. In fact, the accuracy of predictions from transformed data actually increased in moving from the development to the cross-validation sample. A comparison with Table 2 indicates that the average accuracy for prediction in the development sample was higher for the individual module times than it was for the cumulative times. These unusual findings are apparently due to anomalously unpredictable transformed times in the development sample for certain of the modules that were not included in the sample of ten.

Table 7

Percentage of Variance Accounted for by Different Kinds of Predictions: Cumulative Time

Sample	CUM APT	CUM APT Trans.	FADE	CUM APT Trans. Rev.
Dev.	17.6	19.8		
C.-V. Rel.	14.6	22.6	13.8	15.7
C.-V. Abs.	14.5	22.3	13.3	12.5

Table 7 also indicates the accuracies in the cross-validation sample for predictions based on (1) cumulative times that are differentially weighted toward recent performance (these FADE predictions were validated against similarly weighted cumulative times in the cross-validation sample) and (2) reversed transformations. In each case, the accuracies were quite similar to those for untransformed cumulative data.

For a homogeneous series of modules, the reliability of the cumulative scores should increase with the addition of new modules, just as the reliability of a test should increase

with the addition of new items. It was assumed, therefore, that the accuracy of predictions for cumulative scores would also increase as the course progressed.

Table 8 indicates the way in which the accuracies actually changed as the course progressed. There was an overall increase in accuracy, but there were also strong local variations. The first two modules were unusually predictable (relative to other individual modules in the sample). These were followed by a series of modules, many of which were either unusually unpredictable or relatively unrelated to one another. From about Module 10 on, the accuracy of the predictions increased, in the expected manner, to a level considerably above the level found for individual modules.

Table 8

Percentage of Variance Accounted for by Different Kinds of Prediction as a Function of Position in Course: Cumulative Time

	C.-V. Rel.				C.-V. Abs.			
Point in Course (Module)	CUM APT	CUM APT Trans.	FADE	CUM APT Trans Rev.	CUM APT	CUM APT Trans.	FADE	CUM APT Trans. Rev.
1	12.0	19.0	12.1	13.1	11.5	18.8	9.8	11.9
2	14.1	21.8	13.4	15.3	13.9	21.0	13.2	11.1
4	13.4	22.3	12.2	14.1	13.1	22.1	11.8	9.9
6	13.6	21.8	11.7	14.5	13.5	21.6	11.5	10.6
8	13.1	20.9	9.7	13.9	13.0	20.2	9.7	9.9
10	12.3	19.0	8.3	12.8	12.4	18.8	8.2	10.0
12	13.3	21.0	11.6	14.1	13.5	20.0	11.6	11.3
14	13.1	21.4	8.9	14.1	13.2	21.5	6.9	11.4
19	20.1	29.3	24.5	22.0	20.1	29.5	24.2	19.2
20	20.6	29.9	25.8	22.7	20.5	29.6	25.6	19.9

It might be noted that the accuracies of both the reversed transformation and FADE increased relative to the accuracy of CUM APT toward the end of the course. However, it is impossible to tell whether this trend might be expected to continue in a longer course.

Root Mean Square (RMS) Errors for Related Predictions

Since the predictions for untransformed data are additive, the errors for predictions of three related values, all associated with course completion, are identical. One of these is the predicted time for course completion (completion time, or CT), predicted from various points in the course. These predictions are needed to plan various administrative actions associated with course completions. The second is predicted time for the entire course (total time, or TT), again predicted from various points in the course. These

predictions provide the best measure of the student's utility to the Navy, and should be a major factor in selecting attrites. The last is the prediction of discrepancies between an initial prediction of TT based on APT, and the time that will actually be required for the entire course (final discrepancy, or FD). These predictions are used in the current system as a basis for assignments to night school.

Table 9 gives the RMS errors at various points in the course for predictions based on APT, SUM, ADJ, and PAST. The predictions of total time for APT consist of actual cumulative times to the points of prediction plus the predictions of APT CT. For purposes of comparison, errors for CT and TT using reversed transformations of SUM have also been provided.

Table 9

Errors in Predictions of Completion Time, Total Time, and Final Discrepancy as a Function of Position in Course

Point in Course (Module)	RMS Errors (hours)					
	APT	SUM	ADJ	PAST	CT SUM Trans. Rev.	TT SUM Trans. Rev.
1	22.0	19.9	24.2	24.4	20.0	19.9
2	19.3	15.3	20.2	20.4	15.3	15.4
4	17.8	15.2	19.3	19.4	14.5	14.5
6	16.6	13.7	16.7	16.9	13.2	13.2
8	13.6	11.5	13.8	13.9	11.1	11.1
10	12.7	11.1	12.9	12.9	10.7	10.7
12	11.6	9.9	11.3	11.4	9.7	9.7
14	10.0	9.0	10.0	10.2	8.8	8.8
19	2.9	2.8	3.3	3.3	3.0	2.9
20	2.2	2.2	2.6	2.6	2.6	2.4

The errors decline in a fairly linear fashion as the course progresses. Errors for ADJ and PAST are similar to but slightly greater than those for APT, suggesting that the corrections used in the current system may be ill-advised. All of these are higher than those for SUM or the predictions provided by reversed transformations of SUM although the differences are not very large in the latter parts of the course.

Accuracy: Completion Time (CT)

Table 10 indicates both relative and absolute accuracies for predictions of CT based on several indices. The relative accuracies of these predictions were quite stable over most of the course, but dropped off toward the end. The accuracies of all predictions based in part on past performance were also somewhat lower at the beginning of the course. Predictions based on transformed data were clearly the most accurate, and CT APT was clearly the least accurate.

Table 10

Percentage of Variance Accounted for by Different Kinds of Prediction, as a Function of Position in Course: Completion Time

Point in Course (Module)	CT APT	CT SUM	CT ADJ	CT PAST	CT SUM Trans.	CT SUM Trans. Rev.
C.-V. Rel.						
1	20.3	35.3	30.0	29.2	45.3	37.9
2	20.5	50.0	46.2	44.4	57.2	51.3
4	20.5	43.9	39.4	36.5	57.6	47.3
6	21.2	47.5	43.4	39.8	59.9	50.1
8	22.8	46.6	42.6	37.5	56.5	48.5
10	22.8	43.3	39.4	34.7	54.7	45.7
12	23.4	44.4	41.0	35.1	56.7	46.8
14	24.1	39.7	36.0	28.5	51.5	42.5
19	10.0	16.7	14.1	13.6	19.1	15.3
20	7.0	10.5	7.8	7.0	10.0	6.8
C.-V. Abs.						
1	20.2	34.7	3.4	1.4	44.7	34.0
2	20.3	49.9	12.1	11.2	56.9	49.9
4	20.4	41.9	6.4	5.5	57.6	46.9
6	21.1	46.3	19.8	18.2	59.8	49.6
8	22.6	44.8	20.0	18.6	56.5	48.2
10	22.5	41.4	20.1	20.1	53.8	44.5
12	23.1	43.1	26.5	24.5	56.4	45.6
14	24.1	39.3	24.9	21.6	50.5	41.2
19	9.4	10.8	-19.2	-22.6	14.9	1.5
20	7.0	8.3	-22.1	-27.5	6.8	-30.1

The absolute accuracies of all predictors except CT ADJ and CT PAST were essentially the same as the relative accuracies over most of the course. The absolute accuracies of both CT ADJ and CT PAST dropped to a level that, both early and late in the course, was considerably below CT APT. The absolute accuracy of the reversed transformations also dropped to a low level late in the course. In fact, toward the end of the course, all three of the latter predictors were less accurate than were predictions of average time for all students.

<u>Accuracy: Total Time (TT)</u>

Table 11 indicates both relative and absolute accuracies for predictions of TT. All approach perfect predictions at the end of the course. There are sizeable differences in relative accuracies over the first third of the course, but these have been washed out by the last third of the course. The same is true for absolute accuracies, but the differences are somewhat larger. Since the errors are the same as for CT, the order of absolute accuracies is the same.

Table 11

Percentage of Variance Accounted for by Different Kinds of Prediction as a Function of Position in Course: Total Time

Point in Course (Module)	TT APT	TT SUM	TT ADJ	TT PAST	TT SUM Trans.	TT SUM Trans. Rev.
C.-V. Rel.						
1	29.6	42.9	37.9	37.3	52.6	45.0
2	47.7	65.8	63.3	62.0	71.2	66.5
4	56.4	67.4	64.8	63.0	76.7	69.3
6	63.4	73.2	71.2	69.3	80.7	74.5
8	76.3	81.5	80.2	78.2	85.8	81.9
10	79.2	82.7	81.6	80.1	87.2	83.3
12	83.0	85.9	85.2	83.5	89.8	86.4
14	87.0	88.3	87.8	86.1	90.6	88.7
19	98.9	98.9	98.8	98.9	98.7	98.9
20	99.3	99.3	99.2	99.3	98.9	99.3
C.-V. Abs.						
1	29.3	42.2	14.4	12.6	51.8	42.1
2	45.6	65.8	40.0	39.3	71.4	65.5
4	53.7	66.1	45.5	45.0	77.0	69.2
6	59.8	72.6	59.1	58.2	80.9	74.4
8	73.0	80.8	72.1	71.6	85.8	81.9
10	76.4	82.1	75.6	75.6	87.2	83.3
12	80.5	85.5	81.3	80.9	89.6	86.3
14	85.3	88.2	85.4	84.8	91.5	88.7
19	98.8	98.8	98.4	98.4	98.7	98.8
20	99.3	99.3	99.1	99.0	98.9	99.2

Accuracy: Final Discrepancies (FD)

Table 12 indicates both relative and absolute accuracies for predictions of FD. Again, they all approach perfect prediction at the end of the course. Differences in relative accuracies are small throughout the course. Differences in absolute accuracies are much larger, particularly over the first third of the course.

Selection: Recent Changes

The current system provides a display, on the DSPR, of the ratios between actual cumulative time and predicted cumulative time (CUM APT) at the completion of each of the seven most recent modules. It also provides cutting scores to flag differences between the last two ratios or between the first and the last of the seven ratios.

Table 12

Percentage of Variance Accounted for by Different Kinds of Prediction as a Function of Position in Course: Final Discrepancy

Point in Course (Module)	FD APT[a]	FD SUM[a]	FD ADJ	FD PAST	FD SUM Trans.	FD SUM Trans. Rev.
C.-V. Rel.						
1	29.4	29.4	28.9	30.6	33.2	29.1
2	56.6	56.6	55.8	57.5	59.3	57.0
4	58.7	58.7	57.9	58.6	67.0	60.3
6	66.1	66.1	65.4	65.3	72.7	67.0
8	76.4	76.4	75.8	75.9	79.8	76.5
10	78.2	78.2	77.6	77.9	81.8	78.5
12	82.1	82.1	81.7	81.5	85.5	82.4
14	85.2	85.2	84.9	84.3	87.8	85.4
19	98.6	98.6	98.6	98.6	98.2	98.6
20	99.1	99.1	99.1	99.0	98.5	99.1
C.-V. Abs.						
1	10.6	26.8	-8.2	-10.5	31.6	25.0
2	31.2	56.7	24.1	23.3	59.6	55.3
4	41.4	57.2	31.1	30.4	67.3	60.1
6	49.1	65.4	48.2	47.2	72.8	66.9
8	65.9	75.7	64.8	64.1	79.9	76.5
10	70.1	77.4	69.2	69.2	81.7	78.3
12	75.3	81.7	76.4	75.7	85.2	82.3
14	81.4	85.1	81.6	80.8	87.9	85.4
19	98.5	98.5	98.0	98.0	98.1	98.4
20	99.1	99.1	98.8	98.8	98.5	98.9

[a]Values of FD APT and FD SUM are identical for C.-V. Rel.

Table 13 indicates selections based on the difference between two successive ratios of cumulative times. As would be expected, selection decreases as the course progresses, with selections early in the course many times more likely than selections late in the course. The local variations in the curve are due to variations in the length of individual modules. In fact, there was a correlation of .96 between the percentages of students selected at different points in the course and the ratios of mean module times to mean cumulative times at these same points.

Table 13 also indicates selections based on the differences between ratios (using APT) for two successive individual modules. There is considerable variation from point to point within the course, but no systematic change as the course progresses. One difficulty with the use of individual modules is that the cutting score for differences is quite high (.68). An alternative would be to flag changes on the basis of differences between ratios computed over a series of successive modules. When selections were based on differences between ratios for successive groups of three modules each, the cutting score was reduced (.45) but not by much.

Table 13

Percentage Selected on Basis of Differences in Performance at Successive Points in Course

Point in Course (Module)	Ratios (%)	
	Cumulative Times	Successive Module Times
2	40.3	13.8
4	20.3	12.7
6	11.5	14.6
8	20.3	10.4
10	9.0	16.3
12	10.7	13.5
14	7.0	14.6
19	13.0	18.6
20	3.1	20.0

Two forms of bias were investigated for selections based on differences in ratios for both single modules and groups of three modules. As might be expected, selection was inversely related to the size of the ratio at the first of the two points involved. For single modules, the average percentages of students selected were 11.0, 13.2, and 20.6 for the high, middle, and low thirds of the distribution of initial ratios respectively. For ratios based on the sum of three modules, the percentages selected from each third of the distribution were 10.2, 15.9, and 18.9. Selection was also inversely related to predicted time at the second of the two points involved. For single modules, the averge percentages selected were 10.7, 14.3, and 20.0 for the high, middle, and low thirds of the distribution of predicted times respectively. For ratios based on the sum of three modules, the percentages selected from each third of the distribution were 12.0, 13.8, and 19.2.

Selection: Unexpectedly Poor Performance

The primary index of overall performance provided by the present system is the ratio between cumulative time and CUM APT. This, as indicated earlier, appears on the DSPR. It is also used to flag students who exceed predicted time by more than a certain percentage. The more obvious alternatives to such an index are (1) a similar ratio based on FADE, and (2) the difference between cumulative time and APT for cumulative time based on transformed data (CUM APT Trans.).

Another index, which is a function of the final discrepancy predicted from ADJ (FD ADJ), is used for assignments to extra study (night school) and the initiation of Academic Review Boards. The exact nature of this index will be discussed later. However, there are a number of obvious alternatives to FD ADJ; namely, FD APT, FD SUM, FD PAST, and, as a representative of indices based on transformed data, FD SUM Trans. It should be noted that the cutting scores for indices based on predictions of FD from untransformed data, unlike those considered previously, are fixed differences rather than fixed ratios.

Table 14 indicates the way in which selections varied as a function of location in the course. The first four columns provide data for indices related to cumulative time. For CUM APT Trans., separate data have been provided for selections based on a fixed

difference and selections based on k x SEE. The change in FADE values is fairly flat. The remaining values tend to decline as the course progresses, suggesting that the compensations for distributional effects, whether by ratios or transformations, may be too large. It should be noted, though, that the decline is just as pronounced for the cuts based on k x SEE. The drop over the 20 modules in this course was only about 25 percent, but the problem might well be more extreme in a longer course.

Table 14

Percentage Selected by Different Indices of Unexpectedly Poor Performance as a Function of Position in Course: Groups of Modules

Point in Course (Module)	CUM APT	CUM APT Trans. Diff.	CUM APT Trans. k x SEE	FADE	FD APT	FD SUM	FD SUM Trans.	FD ADJ	FD PAST
1	15.0	16.4	16.3	14.4	1.1	4.4	4.4	15.0	15.8
2	16.9	17.8	17.1	16.7	4.7	12.5	11.4	15.3	17.2
4	15.0	15.8	15.7	13.9	9.7	14.4	13.3	15.6	15.6
6	15.3	15.8	15.6	15.6	10.0	15.0	14.4	15.3	15.8
8	14.2	15.8	15.4	15.8	17.5	17.8	16.7	16.4	15.6
10	15.0	14.4	14.9	13.9	19.2	16.9	16.1	16.1	15.0
12	15.8	14.7	14.7	15.3	19.7	18.1	18.3	15.0	15.3
14	15.6	14.4	14.2	14.2	20.0	16.7	17.5	15.3	14.4
19	13.6	12.8	13.6	15.0	24.2	17.2	19.2	13.3	12.8
20	13.6	12.5	13.5	15.3	23.9	16.9	18.6	12.8	12.5

The FD indices fall into two distinct groups. The values for FD ADJ and FD PAST tend to decline, just as do those for the indices related to cumulative time. The remaining values increase from very low levels of selection early in the course to high levels of selection late in the course. The reason for the low early levels is the paucity of new information for predicting the final discrepancy. This is most obviously the case for FD APT. It might be noted, though, that the values for FD SUM and FD SUM Trans. reach fairly high levels of selection by the end of the second module. Nevertheless, individual cutting scores were used for all three of these indices in the remaining analyses.

The selection for mandatory extra study used in the present system is actually based on FD ADJ divided by predicted time to complete the course (CT APT). The premise of this system was to manage the student toward the original prediction for total time. The first column of Table 15 indicates that the likelihood of selections based on this criterion increases considerably as the course progresses. This has been recognized as a problem, so the current system shifts from this criterion to one based on the ratio of cumulative times as soon as 75 percent of the course has been completed. Even within this range, though, there is still considerable variation. There is an almost fourfold increase in selections between the first of the course and the 75 percent point. Selections based on FD ADJ without the division, on the other hand, remain quite constant throughout the course.

Table 15

Percentage Selected for Extra Study, With and Without Adjustment, as a Function of Position in Course

Point in Course (Module)	Not Adjusted for Selection		Adjusted for Selection	
	Current Criterion	FD ADJ	Current Criterion	FD ADJ
1	7.6	14.9	11.5	21.4
2	10.7	15.5	13.0	17.2
4	12.1	15.8	14.9	19.2
6	14.4	15.5	13.5	15.5
8	17.7	16.6	17.5	15.5
10	18.0	16.3	15.8	14.1
12	18.6	15.2	16.6	12.1
14	22.5	15.5	20.0	11.8
19	33.8	13.5	33.8	14.1
20	37.2	13.0	33.5	11.8

The actual bias in selections, however, is probably not this extreme. Since extra study time is not counted against the student in computing these indices, assignments early in the course will tend to reduce the likelihood of assignments late in the course. To assess the effect of such assignments, the system was simulated, assuming that a 2-hour night school would be available at the completion of each of the ten modules in the sample. The results can be found in the last two columns of Table 15. As anticipated, the increasing trend in selections using the current criterion was reduced, and a decreasing trend was engendered for selections based directly on FD ADJ. However, variability was still greater for the current criterion.

A more detailed analysis of the selections made during the simulation indicated that fewer different students would be selected by the current criterion (31%) than by FD ADJ (38%). Once a student is selected by the current criterion, there is a good chance that he will be selected on all subsequent occasions (43%), which is a definite disadvantage. This inability to work one's way out of night school assignments is far less likely with selections based on FD ADJ (16%). Because of its undesirable features, the current criterion was dropped from the remaining analyses.

Table 16 indicates the ways in which selections varied as a function of predicted time. Selections for the two indices of cumulative time were fairly similar in all thirds of the predicted time distribution. Selections based on FADE, however, tended to increase with decreases in predicted time. This tendency was not particularly strong, but a bias against the brighter students is certainly undesirable. Selections for the various indices related to final discrepancy were quite similar to one another. There was a tendency for selections from the lower third of the predicted time distributions to be lower than those from the middle and upper thirds. This is a somewhat desirable form of bias.

Table 16

Percentage Selected by Different Indices of Unexpectedly Poor Performance as a Function of Predicted Time: Groups of Modules

Predicted Time	CUM APT	CUM APT Trans.	FADE	FD APT	FD SUM	FD SUM Trans.	FD ADJ	FD PAST
Long	14.9	15.4	13.5	17.4	17.4	15.8	16.8	16.7
Medium	16.3	17.9	15.4	18.5	18.5	18.1	17.8	18.9
Short	16.3	15.7	19.6	12.8	12.8	14.8	10.9	14.2

Table 17 indicates the estimated overlaps between students selected by different indices. The Time variable, which represents selections based purely on long times (no predictions), is provided as a baseline. Since the technique used to estimate overlaps is not particularly accurate for high levels of overlap, overlaps in the range from 91 percent to 99 percent have been designated as L (for large). Obviously, most of these indices will select pretty much the same students. FADE is the only exception, but even it has moderately high overlaps with the remaining indices. The overlaps with Time are moderately high.

Table 17

Percentage of Overlap of Students Selected by Different Indices of Unexpectedly Poor Performance: Groups of Modules

Basis of Selection	CUM. APT	CUM. APT Trans.	FADE	FD APT	FD SUM	FD SUM Trans.	FD ADJ	FD PAST
Time (long)	71	75	59	78	78	74	79	77
CUM APT		L	79	L	L	L	L	L
CUM APT Trans.			76	L	L	100	100	L
FADE				75	75	76	73	78
FD APT					100	L	100	L
FD SUM						L	100	L
FD SUM Trans.							L	L
FD ADJ								L
FD PAST								

Notes. L indicates overlaps between 91 and 99 percent. These values were computed from the last seven modules in the sample, since relationships early in the series are not representative of an extended series.

Table 18 indicates the estimated percentage of overlap between students selected on different modules. The median variable was calculated over all variables other than Time and FADE (the median range of variation in overlap per point was about 3%). In each case, the overlap decreased with increased separation between modules, but the decrease was greater for FADE than for the other types of selection. The average overlap for FADE was considerably less than it was for other forms of selection.

Table 18

Percentage of Overlap of Students Selected for Unexpectedly Poor Performance on Different Groups of Modules

Basis of Selection	Average Overlap	(Distance (Intervening Modules) Between Modules		
		1	3	5
Time (long)	75	89	81	75
Median[a]	76	86	80	76
FADE	52	72	57	50

[a]For CUM APT, CUM APT Trans., FD APT, FD SUM, FD ADJ, FD PAST, and FD SUM Trans.

The ways in which the total selections were distributed over students were determined to a large degree by the amount of overlap between modules. The distributions for all variables except FADE were quite similar to one another. Most students (about 60%) were never selected. The distributions between one and ten selections were relatively symmetrical, dropping from about 6 percent of the students at the ends to about 2 percent of the students in the middle. For FADE, the percentage never selected was somewhat smaller (about 58%). The distribution between one and ten selections fell off from a frequency of 12 percent for one selection to a frequency of 2 percent for ten selections.

Selection: Unexpectedly Good Performance

The preceding analyses have all focused on negative discrepancies or worsening performance, but the instructor must also select students for incentives. In the present system, incentives are awarded only at the end of the course or, in some cases, at a limited number of widely spaced points within the course.

One of the incentives provides formal recognition to students who complete the course in a time that is at least a certain percentage below the predicted time. A forecast for this incentive is provided on the DSPR by a ratio between cumulative time and APT for cumulative time (CUM APT). Students working at or near a level that will merit recognition can be flagged by means of fixed ratios. An obvious alternative to this index is the difference between transformed cumulative time and APT for cumulative time based on transformed data (CUM APT Trans.). A third index, closely related to the first two though not a strict substitute for either, is the ratio of faded cumulative time to FADE.

The current system can also award a day off if the student completes a certain segment of the course in a time that is at least a certain number of days less than the predicted time. The DSPR indicates whether the student is currently eligible for such an award and how far he is ahead of or behind predicted time, but there is no real forecast of future eligibility. In order to facilitate comparison with the other indices of unexpectedly good performance, it was assumed that time off would be awarded at the end of the course and that the likelihood of time off would be predicted by using FD ADJ.

Table 19 indicates the ways in which selections varied as a function of location in the course. For CUM APT Trans., separate columns are provided for selections based on a fixed difference and selections based on k x SEE. For CUM APT and CUM APT Trans., with selections based on a fixed difference, the percentages of selections declined as the course progressed, just as they did for the selection of unexpectedly poor students. The hypothesis of overcompensation received additional support in this case, since the rate of selection based on k x SEE is relatively level. The decline in percentage of selections for FD ADJ was also moderately large, just as it was for the selection of unexpectedly poor students. There was also some decline for FADE, but it was not as large as for the other indices using fixed cutting scores, and, as might have been expected, it levels off in the second half of the course.

Table 19

Percentage Selected by Five Indices of Unexpectedly Good Performance as a Function of Position in Course: Groups of Modules

Point in Course (Module)	CUM APT	CUM APT Trans. Diff.	CUM APT Trans. k x SEE	FD ADJ	FADE
1	19.4	17.8	15.0	17.2	16.1
2	18.1	17.8	15.3	18.1	17.5
4	15.8	15.0	14.7	16.7	15.5
6	16.7	16.9	16.4	16.1	14.7
8	15.0	14.7	14.7	15.8	15.8
10	14.7	15.0	15.3	13.6	13.6
12	14.7	15.0	15.3	12.8	13.9
14	13.9	13.3	14.7	13.6	13.6
19	11.4	12.2	14.7	12.8	15.3
20	10.8	11.9	13.9	13.3	13.9

Table 20 indicates the way in which selections varied as a function of predicted time. For FADE, the likelihood of selection is roughly the same for each third of the predicted time distribution. For the two indices based on APT, the likelihood of selection is higher for short predicted times than for either of the other two categories. This is probably a good form of bias. For FD ADJ, however, the likelihood of selection is almost two and a half times as great for students with long predicted times as it is for students with short predicted times. This is definitely an undesirable form of bias.

Table 20

Percentage Selected by Four Indices of Unexpectedly Good Performance as a Function of Predicted Time: Groups of Modules

Predicted Time	CUM APT	CUM APT Trans.	FD ADJ	FADE
Long	16.0	14.3	21.3	16.9
Medium	13.6	12.9	14.0	14.7
Short	17.4	19.5	8.7	14.8

Table 21 indicates the estimated overlaps between students selected by means of different indices. The Time variable, which represents selections based purely on short time (no prediction), is provided as a baseline. The overlaps between Time and the two indices based on APT were somewhat higher than those between time and FD ADJ or FADE. There was considerable overlap between the two indices based on APT; the remaining overlaps were moderate.

Table 21

Percentage of Overlap of Students Selected by Different Indices of Unexpectedly Good Performance: Groups of Modules

Basis of Selection	CUM APT	CUM APT Trans.	FD ADJ	FADE
Time (short)	69	74	52	56
CUM APT		92	84	76
CUM APT Trans.			80	75
FD ADJ				72

Note. These values were computed for the last seven modules in the sample, since relationships early in the series are not representative of an extended series.

Table 22 indicates the estimated percentage of overlap between students selected on different modules. In all cases, the overlap decreased with increased separation between modules, but the decrease was greater for FADE than it was for the other types of selection. The average overlap for FADE was considerably less than it was for the other forms of selection.

Table 22

Percentage of Overlap of Students Selected for Unexpectedly Good Performance on Different Groups of Modules

Basis of Selection	Average	Distance (Intervening Modules) Between Modules 1	3	5
Time (short)	81	86	85	83
CUM APT	78	88	82	79
CUM APT Trans.	78	88	81	78
FD ADJ	78	85	81	78
FADE	55	75	62	53

The ways in which the total selections were distributed over students were determined to a large degree by the amount of overlap between modules. The distributions for selections based on FD ADJ and the two versions of CUM APT were quite similar. Most students (about 72%) were never selected. The distribution between one and ten selections was roughly symmetrical, dropping from about 6 percent of the students at the ends to about 1 percent of the students in the middle. For FADE, the percentage never selected was smaller (60%). The distribution between one and ten selections fell off in a negatively accelerated manner, from a frequency of 10 percent for 1 selection to a frequency of 1 percent for 10 selections.

DISCUSSION

Editing

Editing had little effect, one way or another. Editing around a regression line, which should be quite sensitive in detecting errors but which is also quite complicated, demonstrated no real advantage over the much simpler procedure of editing on the extremes. Perhaps the most significant finding was that rather radical editing had no real detrimental effect on predictions.

Transformation

Predictions based on transformed data were, in all cases, much more accurate than were predictions based on untransformed data. Part of their advantage lay in their ability to straighten what appeared to be (when graphed) upwardly concave lines of best fit. A fairly frequent complaint about the current system is that there are many unreasonably short (and occasionally negative) predictions for very bright students. Such predictions are a natural result of fitting straight lines to upwardly concave curves.

The only serious difficulty with predictions from transformed data is their complexity. The current system is built entirely on APT for individual modules. When predictions are required for a group of modules, whether for cumulative times or times for course completion, the predictions for individual modules are simply added to one another. Variations in the pattern of assignment are no problem since, again, the predictions for individual modules are simply added to one another. For predictions based on transformed data, on the other hand, separate predictions would be required at each

point in each pattern of assignments for a cumulative time and a time for course completion.

The difficulty of interpreting predictions from transformed data should probably be viewed as a minor difficulty. The predictions can be used directly for selection, and then, if necessary, the transformation can be reversed to facilitate interpretation. In most cases, the accuracy of predictions based on these reversed transformations is comparable to that of those based on untransformed data. Such predictions will avoid some of the problems, mentioned previously, that result from curvilinear relationships.

The increased accuracy of predictions based on transformed data does not, however, have a major impact on the selection of students. For individual modules, the overlap of selections based on APT and APT Trans. was 88 percent. For cumulative modules, whether used to select unexpectedly good or unexpectedly poor students, it was about 93 to 94 percent. The selections based on transformed data did enjoy some advantange in terms of minimizing the unfortunate bias toward identifying bright students (short predicted times) as unexpectedly poor performers.

Use of Data on Past Performance

In almost all cases, predictions that took the student's past performance into account were substantially more accurate than were those that did not. This finding is not too surprising, and its significance is limited by the fact that in many cases predictions based on APT are dictated by their role in indices used for the allocation of incentives. It is generally felt that a student's ability to win rewards late in the course should not be jeopardized by the fact that he worked unusually hard early in the course. The cases in which predictions based on APT were more accurate than certain predictions based on past performance are interesting, since the current system uses predictions based on ADJ to estimate the time required for course completion and to make assignments to extra study. It appears that predictions based on APT might be better for such purposes.

The differences between alternative predictions based in part on past performance are more significant. The slight superiority of SUM to IND was the result of instability since, on purely logical grounds, IND must be at least as good as SUM. It would probably be possible to demonstrate the superior accuracy of IND by using a larger development sample or by using one of the statistical techniques designed to minimize shrinkage. However, the advantage would probably be quite small, and SUM is much easier to use.

IND and SUM were substantially more accurate than was ADJ, but both would require new equations at each point in each pattern of assignments for both individual module time and time to course completion.

Cutting Scores

In several of the analyses, individual cutting scores were considered as a means for eliminating undesirable variations in the number of students selected from different modules. Individual cutting scores of this kind would be much less convenient than general cutting scores; they would be more difficult to establish and more difficult to adjust. Scores such as k x SEE, however, share some of the advantages of the general scores, since they can be modified by changing a single parameter. Compromises of this kind, based perhaps on linear equations, might be considered for the various indices of unexpectedly poor and unexpectedly good performance should the declining rates of selection found in this study become a more serious problem in longer courses.

Differential Weighting of Recent Performance

Most of the indices considered for the allocation of incentives are based on total past performance. Such indices have some advantages, but for certain purposes, particularly in longer courses, they will carry too much inertia. A student who does poorly early in the course will have great difficulty in overcoming the stigma later in the course, and a student who does well early in the course may continue to coast on his early performance long after he has started to perform poorly. It would be desirable to have an index that is more sensitive to recent changes in performance and to use this as the basis for allocating some incentives.

The index based on FADE was investigated as one possibility for such an index. The predictions were reasonably accurate, selections remained fairly stable throughout the course, and both incentives and disincentives were distributed to a wider range of students.

Display of Recent Performance

The current display of ratios between actual and predicted cumulative times at the completion of each of the last seven modules is influenced far too strongly by the student's position in the course. The indices for changes in performance that are based on differences between these same ratios suffer from the same difficulty. The ratios in the display should be replaced by ratios for individual modules. Indices for changes in performance might also be based on differences between ratios for individual modules, but such ratios are not sufficiently stable for the detection of significant changes in behavior. Consideration should be given to differences between ratios for groups of modules or to techniques that might better reveal a linear trend in a series of ratios.

Time Off

The present procedure of awarding time off on the basis of time saved is intuitively appealing, but it tends to be strongly biased toward rewarding the less gifted students (long predicted times). It would be more equitable and probably more effective if time off were awarded on the basis of time saved that has been adjusted in some way for variations in time predicted. One possibility is the ratio between actual and predicted cumulative times with a minimum cutting score for the difference between the two.

Extra Study

The current technique of assigning extra study is too strongly biased toward selecting students more frequently as the course progresses. There are several obvious alternatives. One, which was analyzed in some detail, is to use FD directly, without dividing by CT. An even simpler technique is to use the ratio of actual to predicted cumulative time throughout the course.

Generalization to Other Courses

There is a good possibility that individual module times in the AFUN course are somewhat less predictable, whether on the basis of APT or times from other modules, than are those in most courses. This might limit the generality of the current findings but an increase in predictability would probably tend to intensify certain effects found in this study. It appears that many of the selections made in this study were the result of gross errors that were large enough to trigger selection by all of the procedures being evaluated, including those that were not based on predictions of any kind. If errors of this

kind were reduced, the overlaps between alternative procedures would be reduced, and biases due to factors such as predicted time and position in course would probably be increased. In a course where there is less overlap and more bias, the practical utility of predictions based on transformed data or on more refined techniques for handling past performance might be greater than they were in this course.

Two aspects of computer-generated reports were not addressed in this evaluation. The first is the use of information on such things as the number of remedial assignments as an aid in the differential diagnosis of student difficulties. Data of the kind needed for such analyses were not readily available when this project was initiated, but they are now. The second in a rational system for dropping students from training. Data from the aviation fundamentals course alone would not have provided an adequate basis for developing and evaluating such a system. Both problems are important and should be pursued.

RECOMMENDATIONS

1. Certain procedures used in the current system could be improved with minimum cost. Among these are those used in displaying performance on recent modules, in selecting students for certain positive incentives, and in selecting students for assignment to extra study. The system should be modified to incorporate these new procedures.

2. Other procedures used in the current system could be improved only through revisions that are fairly extensive, or that might place serious limitations on other aspects of the system. Among these are the use of predictions based on transformed data, new techniques for basing predictions on past performance, and the use of individual rather than general criteria for various types of selection.

GLOSSARY OF ABBREVIATIONS AND ACRONYMS

Abs.	Absolute accuracy of prediction; one minus the square of the ratio of the root mean square error to the standard deviation of the scores.
ADJ	Adjusted prediction of time on single module or series of successive modules; prediction based on aptitude (APT) multiplied by ratio of actual to predicted (APT) times summed over all previously completed modules.
APT	Prediction of time on single module or series of successive modules from multiple linear regression on ASVAB scores, year of birth, and years of education.
ASVAB	Armed Services Vocational Aptitude Battery.
CUM	Cumulative; prediction of cumulative time on a series of successive modules.
C.-V.	Cross-validation; sample used to evaluate indices computed from development (Dev.) sample.
CT	Completion time; predicted time required for course completion from any point in the course--based on any of several kinds of prediction.
Dev.	Development; sample used to evaluate indices computed from development (Dev.) sample.
DSPR	Daily student progress report.
FADE	Predictions of a weighted composite of times on previously completed modules in which weights decrease with increasing distance from current module; from multiple linear regression on same variables as APT.
FD	Final discrepancy; prediction of discrepancy between initial prediction of total time (TT) based on APT and actual total time--made from any point in course and based on any of several kinds of prediction.
IND	Prediction of module times from multiple linear regression on same variables as APT plus individual times for each previously completed module.
k x SEE	The standard error of estimate times a constant; used as a criterion for selection.
PAST	Prediction of time on single module or series of successive module that is made by multiplying average time (all students) by the ratio of individual times to average times (all students) summed over all previously completed modules.
Rel.	Relative accuracy of prediction; one minus the square of the ratio of the standard error of estimate to the standard deviation of the scores.

RMS	Root mean square.
SUM	Prediction of time on single module or series of successive modules from multiple linear regression on same variables as APT plus sum of actual times on all previously completed modules.
Time	Selection of students on basis of actual times (long or short). Used as baseline for selections based on discrepancies from predicted times.
Trans.	Predictions of times following logarithmic transformation.
Trans. Rev.	Predictions of times following logarithmic transformation (Trans.) followed by antilogarithmic transformation of initial predictions.
TT	Total time; predicted time for completing entire course--made from any point in course and based on any of several kinds of prediction (in case of TT APT, prediction is actual time to that point plus CT APT).

DISTRIBUTION LIST

Chief of Naval Material (NMAT 00), (NMAT 05)
Deputy Chief of Naval Material (Technology)
Chief of Naval Education and Training (015), (N-5)
Chief of Naval Technical Training (016)
Commander in Chief U.S. Atlantic Fleet
Commander in Chief U.S. Pacific Fleet
Commander Naval Weapons Center
Commander Training Command, U.S. Atlantic Fleet
Commander Training Command, U.S. Pacific Fleet
Director, Management Information and Instructional Activity Branch Office, Memphis
Director, Training Analysis and Evaluation Group (TAEG)
Commander, Army Research Institute for the Behavioral and Social Sciences, Alexandria (PERI-ASL)
Commander, Air Force Human Resources Laboratory, Brooks Air Force Base (Scientific and Technical Information Office)
Defense Technical Information Center (DDA) (12)

ATE

LMED

-8